CLAIMS

WHAT IS CLAIMED IS:

1. A method for identifying amino acid residues for variation in a protein variant library in order to affect a desired activity, said method comprising:

(a) receiving data characterizing a training set of a protein variant library, wherein protein variants in the library have systematically varied sequences, and

wherein the data provides activity and sequence information for each protein variant in the training set;

(b) from the data, developing a sequence activity model that predicts activity as a function of amino acid residue type and corresponding position in the sequence; and

(c) using the sequence activity model to identify one or more amino acid residues at specific positions in the systematically varied sequences that are to be varied in order to impact the desired activity.

2. The method of claim 1, further comprising:

(d) using the sequence activity model to identify one or more amino acid residues that are to remain fixed in a new protein variant library.

3. The method of claim 1, wherein the protein variant library comprises naturally occurring proteins or proteins derived therefrom.

4. The method of claim 3, wherein the naturally occurring proteins comprise proteins that are encoded by members of a single gene family.

5. The method of claim 1, wherein the protein variant library comprises proteins that are obtained by using a recombination-based diversity generation mechanism.

6. The method of claim 1, further comprising performing DOE to identify the systematically varied sequences.

7. The method of claim 1, wherein the activity is not protein stability.

8. The method of claim 1, wherein the sequence activity model is a regression model.

9. The method of claim 1, wherein the sequence activity model is a partial least squares model.

10. The method of claim 1, wherein the sequence activity model is a neural network.

11. The method of claim 1, wherein using the sequence activity model to identify one or more amino acid residues further comprises identifying sequences for use in a recombination-based diversity generation mechanism, wherein said sequences comprise variations in the one or more amino acid residues identified in (c).

12. The method of claim 1, wherein using the sequence activity model comprises identifying a sequence predicted by the model to have a highest value of the desired activity.

13. The method of 12, wherein using the model further comprises selecting subsequences of the best sequence.

14. The method of claim 1, wherein using the sequence activity model to identify one or more amino acid residues comprises using the sequence activity model to rank residue positions in order of impact on the desired activity.

15. The method of claim 1, wherein using the sequence activity model to identify one or more amino acid residues comprises using the sequence activity model to rank residue types at residue positions in order of impact on the desired activity.

16. The method of claim 1, wherein using the model comprises using the model as a fitness function in a genetic algorithm.

17. The method of claim 1, wherein using the sequence activity model to identify one or more amino acid residues at specific positions in the systematically varied sequences comprises identifying one or more sequences for use in generating a new protein variant library.

18. The method of claim 17, wherein the sequences are oligonucleotide sequences encoding variations of the one or more identified amino acid residues.

19. The method of claim 18, further comprising performing mutagenesis or a recombination-based diversity generation mechanism using the oligonucleotide sequences to generate the new protein variant library.

20. The method of claim 19, wherein performing mutagenesis or a recombination-based diversity generation mechanism is used in a directed evolution procedure.

21. The method of claim 18, wherein the oligonucleotide sequences encode at least a portion of (i) a naturally occurring parent protein having the highest activity among naturally occurring parent proteins, or (ii) a sequence predicted by the sequence activity model to have the highest activity.

22. The method of claim 17, further comprising developing a new sequence activity model using activity and sequence data characterizing the new protein variant library.

23. The method of claim 17, further comprising selecting one or more members of the new protein variant library for production.

24. The method of claim 23, further comprising expressing one or more of the selected members of the new protein variant library.

25. The method of claim 23, further comprising:

(i) providing an expression system from which a selected member of the new protein variant library can be expressed; and

(ii) expressing the selected member of the new protein variant library.

26. The method of claim 1, wherein the one or more amino acid residues identified in (c) are identified in a reference sequence predicted using the sequence activity model or a reference sequence that describes a member of the protein variant library.

27. A method for identifying amino acid residues for variation in a protein variant library in order to affect a desired activity, said method comprising:

(a) receiving data characterizing a training set of a protein variant library comprising proteins that were obtained by performing classical or synthetic DNA shuffling on nucleic acids encoding all or part of one or more naturally

occurring parent proteins, wherein the data provides activity and sequence information for each protein variant in the training set;

(b) from the data, developing a sequence activity model that predicts activity as a function of amino acid residue type and corresponding position in the sequence; and

(c) using the sequence activity model to identify one or more amino acid residues, in proteins of the library, that are to be varied in order to impact the desired activity.

28. A method for identifying amino acid residues for variation in a protein variant library in order to affect a desired activity, said method comprising:

(a) receiving data characterizing a training set of a protein variant library, wherein the data provides activity and sequence information for each protein variant in the training set;

(b) from the data, developing a sequence activity model that predicts activity as a function of amino acid residue type and corresponding position in the sequence; and

(c) using the sequence activity model to identify one or more amino acid residues, in proteins of the protein variant library, that are to be varied in order to identify one or more sequences for use in a directed evolution procedure.

29. The method of claim 28, wherein the sequences are oligonucleotide sequences encoding variations of the one or more identified amino acid residues.

30. A method for identifying amino acid residues for variation in a protein variant library in order to affect a desired activity, said method comprising:

(a) receiving data characterizing a training set of a protein variant library, wherein the data provides activity and sequence information for each protein variant in the training set;

(b) from the data, developing a sequence activity model that predicts activity as a function of amino acid residue type and corresponding position in the sequence;

(c) using the sequence activity model to rank residue positions or residue types at specific residue positions in order of impact on the desired activity;

(d) using the ranking to identify one or more amino acid residues, in proteins of the protein variant library, that are to be varied or fixed in order to impact the desired activity.

31. A method for generating an optimized protein variant library, said method comprising:

(a) receiving data characterizing a training set of a protein variant library,

wherein protein variants in the library have systematically varied sequences, and

wherein the data provides activity and sequence information for each protein variant in the training set;

(b) from the data, developing a sequence activity model that predicts activity as a function of amino acid residue type and corresponding position in the sequence;

(c) using the sequence activity model to select one or more amino acid residues at specific positions in the systematically varied sequences that are predicted to provide desired activity;

(d) generating an optimized protein variant library,

wherein the sequences of the members of the optimized protein variant library each comprise the one or more selected amino acid residues.

32. A computer program product comprising a computer readable medium on which is provided program instructions for identifying amino acid residues for variation in a protein variant library in order to affect a desired activity, said instructions comprising:

(a) code for receiving data characterizing a training set of a protein variant library,

wherein protein variants in the library have systematically varied sequences, and

wherein the data provides activity and sequence information for each protein variant in the training set;

(b) code for using the data to develop a sequence activity model that predicts activity as a function of amino acid residue type and corresponding position in the sequence; and

(c) code for using the sequence activity model to identify one or more amino acid residues at specific positions in the systematically varied sequences that are to be varied in order to impact the desired activity.

33. The computer program product of claim 32, wherein the program instructions further comprise:

(d) code for using the sequence activity model to identify one or more amino acid residues that are to remain fixed in a new protein variant library.

34. The computer program product of claim 32, wherein the program instructions further comprise code for performing DOE to identify the systematically varied sequences.

35. The computer program product of claim 32, wherein the sequence activity model is a regression model.

36. The computer program product of claim 32, wherein the sequence activity model is a partial least squares model.

37. The computer program product of claim 32, wherein the sequence activity model is a neural network.

38. The computer program product of claim 32, wherein the code for using the sequence activity model comprises code for identifying a sequence predicted by the model to have a highest value of the desired activity.

39. The computer program product of 38, wherein the code for using the model further comprises code for selecting subsequences of the best sequence.

40. The computer program product of claim 32, wherein the code for using the sequence activity model to identify one or more amino acid residues comprises code for using the sequence activity model to rank residue positions in order of impact on the desired activity.

41. The computer program product of claim 32, wherein the code for using the sequence activity model to identify one or more amino acid residues comprises code for using the sequence activity model to rank residue types at residue positions in order of impact on the desired activity.

42. The computer program product of claim 32, wherein the code for using the model comprises code for using the model as a fitness function in a genetic algorithm.

43. The computer program product of claim 32, wherein the code for using the sequence activity model to identify one or more amino acid residues at specific

positions in the systematically varied sequences comprises code for identifying one or more sequences for use in generating a new protein variant library.

44. The computer program product of claim 43, wherein the sequences are oligonucleotide sequences encoding variations of the one or more identified amino acid residues.

45. The computer program product of claim 44, wherein the oligonucleotide sequences encode at least a portion of (i) a naturally occurring parent protein having the highest activity among naturally occurring parent proteins, or (ii) a sequence predicted by the sequence activity model to have the highest activity.

46. The computer program product of claim 43, further comprising code for developing a new sequence activity model using activity and sequence data characterizing the new protein variant library.

47. The computer program of claim 43, further comprising code for selecting one or more members of the new protein variant library for production.

48. The computer program product of claim 32, wherein the code in (c) identifies the one or more amino acid residues in (i) a reference sequence predicted using the sequence activity model or (ii) a reference sequence that describes a member of the protein variant library.

49. A computer program product comprising a computer readable medium on which is provided program instructions for identifying amino acid residues for variation in a protein variant library in order to affect a desired activity, said program instructions comprising:

(a) code for receiving data characterizing a training set of a protein variant library comprising proteins that were obtained by performing classical or synthetic DNA shuffling on nucleic acids encoding all or part of one or more naturally occurring parent proteins, wherein the data provides activity and sequence information for each protein variant in the training set;

(b) code for using the data to develop a sequence activity model that predicts activity as a function of amino acid residue type and corresponding position in the sequence; and

(c) code for using the sequence activity model to identify one or more amino acid residues, in proteins of the library, that are to be varied in order to impact the desired activity.

50. A computer program product comprising a machine readable medium on which is provided program instructions for identifying amino acid residues for variation in a protein variant library in order to affect a desired activity, said program instructions comprising:

(a) code for receiving data characterizing a training set of a protein variant library, wherein the data provides activity and sequence information for each protein variant in the training set;

(b) code for using the data to develop a sequence activity model that predicts activity as a function of amino acid residue type and corresponding position in the sequence; and

(c) code for using the sequence activity model to identify one or more amino acid residues, in proteins of the protein variant library, that are to be varied in order to identify one or more sequences for use in a directed evolution procedure.

51. A computer program product comprising a machine readable medium on which is provided program instructions for identifying amino acid residues for variation in a protein variant library in order to affect a desired activity, said program instructions comprising:

(a) code for receiving data characterizing a training set of a protein variant library, wherein the data provides activity and sequence information for each protein variant in the training set;

(b) code for using the data to develop a sequence activity model that predicts activity as a function of amino acid residue type and corresponding position in the sequence;

(c) code for using the sequence activity model to rank residue positions or residue types at specific residue positions in order of impact on the desired activity;

(d) code for using the ranking to identify one or more amino acid residues, in proteins of the protein variant library, that are to be varied or fixed in order to impact the desired activity.

52. A computer program product comprising a machine readable medium on which is provided program instructions for generating an optimized protein variant library, said program instructions comprising:

(a) code for receiving data characterizing a training set of a protein variant library,

wherein protein variants in the library have systematically varied sequences, and

wherein the data provides activity and sequence information for each protein variant in the training set;

(b) code for using the data to develop a sequence activity model that predicts activity as a function of amino acid residue type and corresponding position in the sequence;

(c) code for using the sequence activity model to select one or more amino acid residues at specific positions in the systematically varied sequences that are predicted to provide desired activity;

(d) code for defining an optimized protein variant library,

wherein the sequences of the members of the optimized protein variant library each comprise the one or more selected amino acid residues.

53. A method of identifying members of a population of biopolymer sequence variants most suitable for artificial evolution, the method comprising:

(a) selecting or screening the members of a population of biopolymer sequence variants for two or more desired objectives to produce a multi-objective fitness data set;

(b) identifying a Pareto front in the multi-objective fitness data set; and,

(c) selecting one or more members proximal to the Pareto front, thereby identifying the members of the population of biopolymer sequence variants most suitable for artificial evolution.

54. The method of claim 53, wherein step (c) comprises:

(i) calculating a weighted sum of the two or more desired objectives for at least some of the members proximal to the Pareto front; and

(ii) selecting at least one member comprising a higher weighted sum than other members proximal to the Pareto front.

55. The method of claim 53, wherein step (c) comprises:

(i) ranking the one or more members according to relative proximity to the Pareto front and relative isolation in sequence space; and,

(ii) selecting at least one member that ranks higher than other members proximal to the Pareto front.

56. A computer program product comprising a computer readable medium having one or more logic instructions for

(a) applying one or more multi-objective evolutionary algorithms to at least one parental biopolymer sequence to produce a set of biopolymer sequence variants;

(b) selecting or screening the members of the set of biopolymer sequence variants for two or more desired objectives;

(c) plotting the set of biopolymer sequence variants as a function of the two or more desired objectives to produce a biopolymer sequence variant plot; and,

(d) identifying a Pareto front in the biopolymer sequence variant plot to identify the members of the set of biopolymer sequence variants comprising multiple improved objectives relative to other members of the set of biopolymer sequence variants.

57. A method of predicting sequences that comprise desired properties, the method comprising:

(a) evolving at least one parental sequence using at least one artificial evolution procedure to produce at least one population of artificially evolved sequences;

(b) selecting or screening the population of artificially evolved sequences for at least one desired property to produce a population of selected artificially evolved sequences;

(c) training a neural network with the population of selected artificially evolved sequences to produce a trained neural network; and,

(d) predicting one or more sequences that comprise the at least one desired property using the trained neural network.

58. A computer system for predicting sequences that comprise desired properties, comprising:

(a) at least one computer system comprising a neural network and a database capable of storing sequences; and,

(b) system software comprising one or more logic instructions for:

(i) evolving at least one parental sequence using at least one artificial evolution procedure to produce at least one population of artificially evolved sequences;

(ii) selecting or screening the population of artificially evolved sequences for at least one desired property to produce a population of selected artificially evolved sequences;

(iii) training the neural network with the population of selected artificially evolved sequences to produce a trained neural network; and

(iv) predicting one or more sequences that comprise the at least one desired property using the trained neural network.

59. A computer program product for predicting sequences that comprise desired properties, comprising a computer readable medium having one or more logic instructions for:

(a) evolving at least one parental sequence using at least one artificial evolution procedure to produce at least one population of artificially evolved sequences;

(b) selecting or screening the population of artificially evolved sequences for at least one desired property to produce a population of selected artificially evolved sequences;

(c) training a neural network with the population of selected artificially evolved sequences to produce a trained neural network; and,

(d) predicting one or more sequences that comprise the at least one desired property using the trained neural network.

60. A method of predicting at least one property of at least one target polypeptide sequence, the method comprising:

(a) identifying one or more motifs common to two or more members of a population of polypeptide sequence variants, wherein at least a subset of the population of polypeptide sequence variants comprises the at least one property, to produce a motif data set;

(b) correlating at least one motif from the motif data set with the at least one property to produce a motif scoring function; and,

(c) scoring the at least one target polypeptide sequence using the motif scoring function, thereby predicting the at least one property of the at least one target polypeptide sequence.

61. A system for predicting at least one property of at least one target polypeptide sequence, comprising:

(a) at least one computer comprising a database capable of storing sequences; and,

(b) system software comprising one or more logic instructions for:

(i) identifying one or more motifs common to two or more members of a population of polypeptide sequence variants, wherein at least a subset of the

population of polypeptide sequence variants comprises the at least one property, to produce a motif data set;

(ii) correlating at least one motif from the motif data set with the at least one property to produce a motif scoring function; and

(iii) scoring the at least one target polypeptide sequence using the motif scoring function to predict the at least one property of the at least one target polypeptide sequence.

62. A computer program product for predicting at least one property of at least one target polypeptide sequence, comprising a computer readable medium having one or more logic instructions for:

(a) identifying one or more motifs common to two or more members of a population of polypeptide sequence variants, wherein at least a subset of the population of polypeptide sequence variants comprises the at least one property, to produce a motif data set;

(b) correlating at least one motif from the motif data set with the at least one property to produce a motif scoring function; and,

(c) scoring the at least one target polypeptide sequence using the motif scoring function to predict the at least one property of the at least one target polypeptide sequence.

63. A system for predicting sequence activities, comprising:

(a) at least one computer comprising a database capable of storing sequences; and,

(b) system software comprising one or more logic instructions for:

(i) selecting a set of parental sequences for at least one activity to produce a set of selected parental sequences;

(ii) subjecting the set of selected parental sequences to one or more artificial evolution procedures to produce a set of evolved sequences;

(iii) selecting the set of evolved sequences for the at least one activity to produce a set of selected evolved sequences;

(iv) providing a sequence-activity plot for the set of sequence variants; and

(v) predicting at least one activity of one or more sequences from the sequence-activity plot.

64. A computer program product for predicting sequence activities, comprising a computer readable medium having one or more logic instructions for:

(a) selecting a set of parental sequences for at least one activity to produce a set of selected parental sequences;

(b) subjecting the set of selected parental sequences to one or more artificial evolution procedures to produce a set of evolved sequences;

(c) selecting the set of evolved sequences for the at least one activity to produce a set of selected evolved sequences;

(d) providing a sequence-activity plot for the set of sequence variants; and,

(e) predicting at least one activity of one or more sequences from the sequence-activity plot.

65. A method of producing libraries of desired sizes, the method comprising:

(a) identifying one or more homologues of at least one initial polypeptide sequence;

(b) comparing the sequences of the homologue(s) and the initial polypeptide;

(c) identifying variable amino acid residues, wherein variable amino acid residues differ with respect to residue type at corresponding positions in the sequences of the homologue(s) and the initial polypeptide sequence;

(d) identifying a set of evolutionarily conserved variable amino acid residues; and

(e) generating a library of protein variants incorporating the set of evolutionarily conserved variable amino acid residues.

66. The method of claim 65, wherein step (b) comprises using at least one substitution matrix to identify the set of evolutionarily conserved variable amino acid residues.

67. The method of claim 65, wherein the library produced by the method comprises a high average fitness as compared to the fitness of the initial polypeptide sequence.

68. The method of claim 65, wherein the homologues comprise a phylogenetic family of polypeptides.

69. The method of claim 65, further comprising screening or selecting members of the library provided in step (e) for one or more desired properties.

70. The method of claim 65, further comprising repeating steps (a)-(e) using at least one screened or selected member as the at least one initial polypeptide in a repeated step (a).

71. A system for producing libraries of desired sizes, comprising:

(a) at least one computer comprising a database capable of storing sets of polypeptide sequences; and,

(b) system software comprising one or more logic instructions for:

(i) identifying one or more homologues of at least one initial polypeptide sequence from a selected evolutionary timescale;

(ii) comparing the sequences of the homologue(s) and the initial polypeptide;

(iii) identifying variable amino acid residues, wherein variable amino acid residues differ with respect to residue type at corresponding positions in the sequences of the homologue(s) and the initial polypeptide sequence; and

(iv) identifying a set of evolutionarily conserved variable amino acid residues.

73. The system of claim 72 wherein the system software further comprises logic instructions for:

(v) identifying a set of oligonucleotide sequences that collectively encode polypeptide variants of the initial polypeptide, wherein the set comprises oligonucleotides that encode the set of evolutionarily conserved variable amino acid residues.

74. A computer program product for producing libraries of desired sizes, comprising a computer readable medium having one or more logic instructions for:

(i) identifying one or more homologues of at least one initial polypeptide sequence/sequence from a selected evolutionary timescale;

(ii) comparing the sequences of the homologue(s) and the initial polypeptide;

(iii) identifying variable amino acid residues, wherein variable amino acid residues differ with respect to residue type at corresponding positions in the sequences of the homologue(s) and the initial polypeptide sequence; and

(iv) identifying a set of evolutionarily conserved variable amino acid residues.

75. The method of claim 1, wherein developing the sequence activity model comprises applying principal component regression to the activity and sequence information.

76. The method of claim 1, wherein developing the sequence activity model comprises using a support vector machine with the activity and sequence information.

77. A method for identifying nucleotides for variation in nucleic acids encoding a protein variant library in order to affect a desired activity, said method comprising:

(a) receiving data characterizing a training set of a protein variant library, wherein the data provides activity and nucleotide sequence information for each protein variant in the training set;

(b) from the data, developing a sequence activity model that predicts activity as a function of nucleotide types and corresponding position in the nucleotide sequence;

(c) using the sequence activity model to rank positions in a nucleotide sequence and/or nucleotide types at specific positions in the nucleotide sequence in order of impact on the desired activity;

(d) using the ranking to identify one or more nucleotides, in the nucleotide sequence, that are to be varied or fixed in order to impact the desired activity.

78. The method of claim 77, wherein the nucleotides to be varied are codons encoding particular amino acids.

79. The method of claim 78, wherein the activity is a function of expression of nucleic acids.

80. A computer program product comprising a machine readable medium on which is provided program instructions for identifying nucleotides for variation in nucleic acids encoding a protein variant library in order to affect a desired activity, said instructions comprising:

(a) code for receiving data characterizing a training set of a protein variant library, wherein the data provides activity and nucleotide sequence information for each protein variant in the training set;

(b) code for developing a sequence activity model from the data, which sequence activity model predicts activity as a function of nucleotide types and corresponding position in the nucleotide sequence;

(c) code for using the sequence activity model to rank positions in a nucleotide sequence and/or nucleotide types at specific positions in the nucleotide sequence in order of impact on the desired activity;

(d) code for using the ranking to identify one or more nucleotides, in the nucleotide sequence, that are to be varied or fixed in order to impact the desired activity.

81. The computer program product of claim 80, wherein the nucleotides to be varied are codons encoding particular amino acids.

82. The computer program product of claim 80, wherein the activity is a function of expression of nucleic acids.

83. A method of defining a library of biological molecules, the method comprising:

(a) receiving an original set of data points representing activity and sequence of multiple biological molecules in a training set;

(b) constructing a bootstrap set of data points selected, with replacement, from the original set of data points;

(c) generating a model from the bootstrap set, which model comprises indicators of the relative importance of individual residues or other units in biological molecules represented by the data points in the bootstrap set;

(d) repeating (b) and (c) multiple times to generate multiple values of each indicator from the model generated in (c);

(e) for each indicator, determining (i) an average or mean value of the multiple values and (ii) a statistical indication of the distribution of the multiple values;

(f) ranking the individual residues or other units on basis of their respective values of (i) and (ii) determined in (e); and

(g) toggling particular ones of the individual residues or other units based on rankings produced in (f) to thereby define the library of biological molecules.

84. The method of claim 83, wherein the original set of data points is generated by systematic variation of a starting sequence.

85. The method of claim 83, wherein (b) comprises constructing the bootstrap set with two or more occurrences of a data point representing the same biological molecule.

86. The method of claim 83, wherein (b) comprises constructing the bootstrap set with no occurrence of a data point having a particular residue or other unit found in at least one of the multiple biological molecules in the training set.

87. The method of claim 83, wherein (c) comprises using a regression technique to generate the model.

88. The method of claim 87, wherein the regression technique is PLS or PCR.

89. The method of claim 83, wherein (b) and (c) are repeated at least about 100 times.

90. The method of claim 83, further comprising generating a p-value for each indicator, wherein the p-value is generated from a mean and standard deviation of multiple values for each indicator.

91. The method of claim 90, wherein (f) comprises ranking the individual residues or other units on the basis of the p-values.

92. The method of claim 83, wherein the biological molecules are proteins.

93. The method of claim 83, wherein the individual residues or other units in biological molecules are amino acids.

94. The method of claim 83, wherein the individual residues or other units in biological molecules are codons for encoding particular amino acids.

95. A computer program product comprising a machine readable medium on which is provided program instructions for defining a library of biological molecules, the instructions comprising:

(a) code for receiving an original set of data points representing activity and sequence of multiple biological molecules in a training set;

(b) code for constructing a bootstrap set of data points selected, with replacement, from the original set of data points;

(c) code for generating a model from the bootstrap set, which model comprises indicators of the relative importance of individual residues or other units in biological molecules represented by the data points in the bootstrap set;

(d) code for repeating (b) and (c) multiple times to generate multiple values of each indicator from the model generated in (c);

(e) code for determining, for each indicator, (i) an average or mean value of the multiple values and (ii) a statistical indication of the distribution of the multiple values;

(f) code for ranking the individual residues or other units on basis of their respective values of (i) and (ii) determined in (e); and

(g) code for toggling particular ones of the individual residues or other units based on rankings produced in (f) to thereby define the library of biological molecules.

96. The computer program product of claim 95, wherein (b) comprises code for constructing the bootstrap set with two or more occurrences of a data point representing the same biological molecule.

97. The computer program product of claim 95, wherein (b) comprises code for constructing the bootstrap set with no occurrence of a data point having a particular residue or other unit found in at least one of the multiple biological molecules in the training set.

98. The computer program product of claim 95, wherein (c) comprises code for using a regression technique to generate the model.

99. The computer program product of claim 98, wherein the regression technique is PLS or PCR.

100. The computer program product of claim 95, further comprising code for generating a p-value for each indicator, wherein the p-value is generated from a mean and standard deviation of multiple values for each indicator.

101. The computer program product of claim 100, wherein (f) comprises code for ranking the individual residues or other units on the basis of the p-values.